Starlight Initiative: Illuminating the Unseen

Bhavya Gupta, Ananya Kashyap, Tanush Goel, Sebu Eisaian
Hungry Hungry Hippo

Department of Information and Computer Sciences, University of California, Irvine

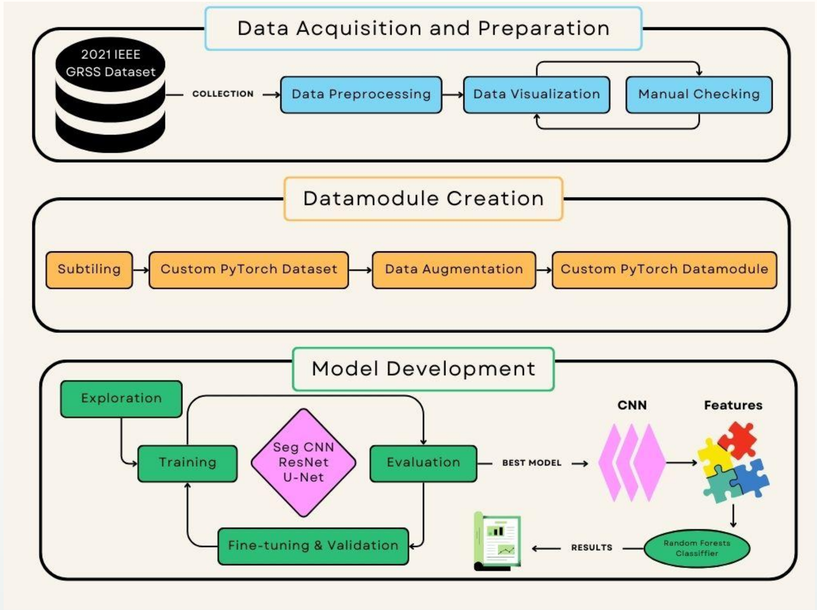
UCI University of California, Irvine

Challenge & Opportunity

Introduction

The stark challenge faced by millions in Africa, living without electricity, underscores a deep-seated issue of access to basic necessities and the resulting community stagnation. This scenario limits crucial opportunities for education, commerce, and global connectivity, further entrenching inequality. However, this challenge also presents a unique opportunity, especially for those with technological access and a passion for software development. Innovations in machine learning and data processing have opened new avenues for identifying non-electrified areas accurately, using high-resolution satellite imagery and advanced modeling techniques like Convolutional Neural Networks (CNNs). By leveraging these technological advancements, developers and policymakers have a powerful tool at their disposal to inform resource allocation, plan infrastructure development, and support humanitarian aid efforts. This dual narrative of challenge and opportunity forms the foundation of initiatives aimed at bridging the electrification gap in Africa, turning data into actionable insights for tangible societal benefit.

Machine Learning Pipeline



Data

Our analysis utilized the IEEE GRSS 2021 ESD dataset, comprising high-resolution Sentinel-2 multispectral imagery, among other satellite data. This dataset is provided in TIFF format, encapsulating a rich array of spectral information across 12 channels in the visible, near-infrared (VNIR), and shortwave infrared (SWIR) spectra, excluding the cirrus band for its lack of ground information.

Data Processing

The data was properly structured and explored to identify outliers and distribution patterns across various spectral bands. Noise reduction using Gaussian filtering, outlier removal using quantile clipping, and data normalization were applied to the different satellite data inputs. Additionally, we explored projecting the VIIRS nighttime data to highlight the brightest areas across time. The data was then ready for further analysis and visualization to provide more in-depth insights into the satellite imagery.

Baseline Models

Our baseline models involved a convolutional neural network, FCN residual neural network, and U-Net. The deep learning model we ultimately chose was the CNN model with random forests to ensure a more robust and accurate model. Therefore, this directly correlated to the baseline CNN which uses learnable filters to extract features from images, hierarchically combining and processing them.

Deep Learning Model

As we chose CNN + random forests, we needed to optimize the hyperparameters. We used a learning rate of 0.0001584, depth of 2, batch size of 31, and 18 epochs to attain the best results. After training this model, we can save it and run random forests on top, which uses ensemble learning to construct a multitude of decision trees to introduce randomness via random feature selection - improving accuracy.

Outcome & Results

Baseline Models w/ Hyperparameter Tuning

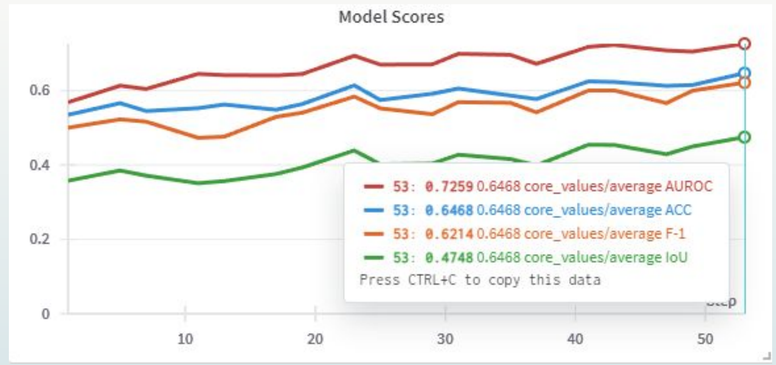
Convolutional Neural Network (CNN): 64.68% accuracy and an AUROC score of 0.7259
FCN Residual Neural Network (FCN ResNet): 61.91% accuracy and an AUROC score of 0.6571
U-Net: 56.38% accuracy and an AUROC score of 0.6573

Deep Learning Model

CNN + Random Forests: 53.29% accuracy with 5000 estimators, log loss, 4 max features, and 500 samples per leaf

Discussion of Results

CNN with random forests performed ~10% worse than without it. We believe this to be the case due to lack of time and processing power needed to ensure random forests can run a very large decision tree. Another factor could be our optimization of validation accuracy while training the CNN model rather than optimizing for validation loss. It is possible that the Bayes algorithm utilized found out that the easiest way to increase the accuracy was to lower training loss in the short term; since we could not have a more long-term solution due to compute power, this approach could have affected us negatively.



Future Work

Next Steps

Broadening Data Integration: Incorporating Sentinel-1 SAR, Landsat-8 multispectral and thermal imagery, and VIIRS nighttime data could provide a more holistic view of environmental conditions. Combining these data sources promises to bolster environmental and land use analyses by offering varied perspectives of Earth's surface.

Model Selection Reevaluation: Our initial choice of models, including Segmentation CNN, FCN ResNet, and U-Net, may benefit from reassessment. Exploring a wider range of models, from simple to complex, could yield insights into optimal modeling approaches and ensure a robust analysis foundation.

Addressing Computational Constraints: Limited computing power restricted our exploration of complex models and thorough parameter tuning. Greater investment in computational resources or the use of cloud computing could facilitate a deeper dive into model potential, enhancing performance and insights.

Time Constraints: One of the biggest limiters for model training is the lack of time as this project was on a short deadline due to being a class. With a longer project duration, more time can be devoted to exploring models, tuning hyperparameters, and training these large models.